Decision Support in Customer Relationship Management Using Data Mining Techniques

Deeksha Bhardwaj
Department of Computer Engineering
G.H. Raisoni Institute of Engineering and Technology, Pune, Maharashtra, India

Dhruv Pandya
Department of Computer Engineering
G.H. Raisoni Institute of Engineering And Technology, Pune, Maharashtra, India

Darshan Patel
Department of Computer Engineering
G.H. Raisoni Institute of Engineering and Technology, Pune, Maharashtra, India

Abstract-With a unbridled increase in international and domestic forms of business, Customer Relationship Management (CRM) has become one of the matters of concern to the enterprise and the entrepreneurs. CRM takes customer as the center and it enchants a new life to the organization system and optimizes its business process increasing its profitability. In order to help enterprises understand the "Product Purchasing Psychology (PPP)" and ways to retain the valued customers we propose data mining techniques. Clustering of customers provide in depth knowledge of their behavior. Clustering is one of the most useful and traditional technique used in data mining. The scope of this paper is to understand and predict the behaviors of the customer with behavior segmentation methodology. The result of the analysis results into enhancing of the customer support and targeting sales of the right product to the customers with better concentration on campaigning product promotion. The policy holders claim dataset for the health insurance company is taken for consideration. This behavior segmentation methodology with clustering is applied here to predict distinct customer segments which help in the production of customized products which takes care of the priorities and preferences of the customers. Apriori association rule which is performed on clusters of claim dataset gives the association amongst the attributes. It is derived from Clustering Based Association Rule Mining (CBARM) model. Association rule is applied on claim dataset which predicts the claim cost and association amongst the attributes that influences the claim cost of the policy holder.

Keywords –Customer Relationship Management (CRM), Public Purchasing Psychology (PPP), Clustering Association Based Association Rule Mining (CBARM).

I. INTRODUCTION

With a Rampant competition in international and domestic business the customer relationship management has become one of the matters of concern for the enterprise and entrepreneurs. CRM can be defined as the method for predicting the customer behavior and selecting appropriate behavior to benefit the company. This concept has been given new lease of life because of the development of the Internet and E-Business. Due to CRM the customer satisfaction also increases and helps in marketing strategy. But one of the important issues in Customer Relationship Management is the customer segmentation and prediction by which the company classifies its customer into pre-defined groups with similar behavioral patterns. Usually company builds a prediction model to find the prospects for the similar products. As a rising subject data mining is playing an important role in decision support models of every walk of life. Data mining uses sophisticated statistical processing of data or artificial intelligence algorithms for discovery of useful trends and patterns from the raw data extracted so that it can yield important insights including prediction model and association rules by which the company can understand the customer in a better way. Customer classification and prediction is the base of implementing CRM. It's the pre-condition to analyze the customer's pattern of consumption and the premise of the personalized marketing and services.

II.PLATFORM BASICS

A) Understanding The Clustering Phenomenon:-

Clustering can be defined as the process of grouping the similar type of physical or abstract objects into classes. Clustering can also be defined as unsupervised classification, because the classification is not dictated/ordered by given class labels. There are many clustering approaches, all based on the principle of maximizing the similarity between objects in the same classes(intra class similarity) and minimizing the similarity between objects of different classes(inter class similarity).this chapter identified extensive work on customer segmentation and data mining techniques and elaborates as follows:

Samira etal. (2007) applied segmentation of customers of trade promotion organization of Iran using a proposed distance function which measures dissimilarities among export basket of different countries based on association rules of concepts. Later, in order to suggest the best strategy for promoting each segment, each cluster is analyzed using RFM model. Variables used for segmentation criteria are “The Value of Group Commodities”, “The Type of Group Commodities” and “the correlation between export group commodities”.

Huang, Chang, and Wu (2009) applied K-means method, Fuzzy C-means clustering method and bagged clustering algorithm to analyze customer value for hunting store in Taiwan and finally concluded that bagged clustering algorithm outcomes the two other method.

Pramod et al. (2011) elaborates the use of clustering to segment customer profiles of a retail store .The study concluded that the k means clustering allows retailers to increase customer understanding and make knowledge – driven decisions in order to provide personalized and efficient customer service.

The main theme of our paper issues clustering technique which predicts the customer behavior for the claim dataset of health insurance claims and the health risks.

III. STRATEGY FOR SEGMENTATION

The main aim of the customer segmentation is indentifying and achieving the profitable sectors and provides products and services that are customer’s common need. Sophist acted customer segmentation give the companies the ability to target more profitable customers, understand the customer’s demands, allocation of the resources and to compete with the rivals. The proposed segmentation framework in given figure 1.

The Main Three Phases of the framework proposed are as follows:

1. Data Preparation Phase
2. Data Clustering Phase
3. Customer Preferences Analyses

A part of the first phase includes collection of data from data store and subsequent data cleansing. Second phase generates Behavioral Segmentation based clusters and profiles of the clusters. Third phase is concerned with identification of customers preferences over products and the risks levels for disease diagnosed, treated and subsequent process of clam settlement.

1. Data Preparation :-

The Database source is the Insurance regulatory Authority Of India‘s Database for health insurance claim dataset .The dataset has 16000 customer claim information for period of one year 2011-2012. The initial dataset contains 36 attributes. The attributes relevant are only taken for analysis. After preprocessing only 21 attributes relevant for analysis are taken into consideration. Health insurance dataset has policy holder whose type of covers is one among the four types such as individual, individual floater, group and group floater.

2. Behavior Analysis:-

In order to describe customer behavior, several attributes can be identified format the policy holders transactions. The policy holders can be characterized by kind of policies opted and with what kind of products the avail. Additionally usage based factors such as how many customers use different policies, during which occasions and how much they tend to avail claim for the health insurance policy opted. The risk associated with each policy holder decides the claim cost. The risk of the policy holders depends on healthy behavior. The diseases diagnosed. Medical history, treatment undergone for recovery process influences the total claim of the policy.

3. Clustering Strategy:-

Customers are divided by use of the cluster analysis. We then form the clusters for the customers divided. The component scores are fed as an input to a cluster model which then assess the similarities between the records/customers and suggests an efficient way of grouping them. Customers claim information’s is based on health risk. The cluster is then formed in such a manner as to reveal significant areas of risk within the insurance profile. The framework of the behavioral is used to predict the expected claim costs for different health risks diagnosed and treatment undergone. K-means is one of the unsupervised learning algorithms that solve the well known clustering problem. The procedure to classify the given dataset through a certain number of clusters is through A priori. The main o objective is to be to define the k centroids, one for each cluster. The cencroids should be placed in a cunning way because different locations cause different results.

4. Association rule based analysis:-

The Analysis of claims data in our health insurance claim dataset is performed using Apriori association algorithm. Associations rule based analysis helps in identifying and establishing claiming patterned and relationships which represents appropriate Utilizations of services.

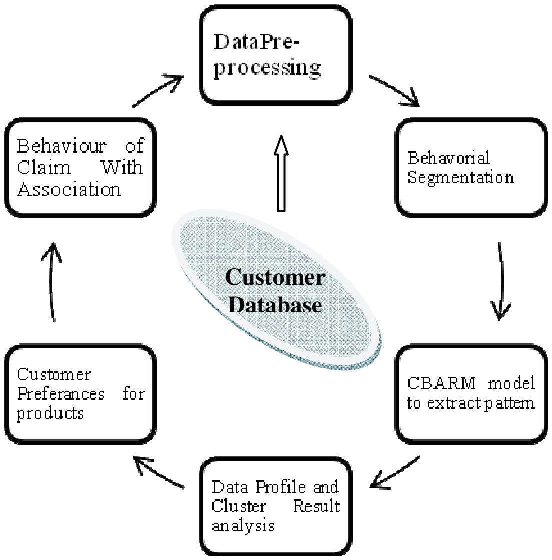
DataPre-processing
Behavorial Segmentation
CBARM model to extract pattern
Data Profile and Cluster Result analysis
Customer Preferances for products
Behaviour of Claim With Association
Customer Database

Figure1. Classification Framework for Different Techniques in data mining in CRM.

IV. EXPERIMENTAL METHODOLOGY AND RESULTS

Clustering an unsupervised learning technique is applied to form cluster of customers based on claim behavior of health insurance customer. The health risk of the customer has significant impact on insurance claim. The WEKA (“Waikato Environment for knowledge analysis”) machine learning work bench 3.7.5 has been considered for the purpose of analysis and test result.

A. Clusters for analyzing customer data.

The formation of the clusters and evaluation under the classes to clusters mode in WEKA is implemented for policy type class attribute. Each cluster shows ser a peculiar type of behavior in customers, from which conclusions are drawn. WEKAfirst ignores the class attribute and generates the clustering. Then during test phase, it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. The total numbers of instances were not fairly distributed if the default seed size is 10 and number of clusters is set to four. It means the sum of the whole clusters was 90%. The distributions of instances were significantly improved, when the seed number was set 100. When the seed size is 10, the number of iterations and sum of squared errors were 10 and 22193.62 respectively. However, if the seed size is set to 100, the number of iteration was 10 and sum of squared errors was 14864.65.

Cluster A- This group is mostly dominated by policy class of group policy holders who may be working in a company or members of co-operative society and so on. The policy holders availed the high claim value compared to others group. Female gender is dominated is the group and claims comes are being mostly diagnosed with pregnancy, childbirth and the puerperium related. Nearly 30% of total claims come from medicines charges. Surgical, consultation and investigation claims are highest among the cluster groups.

Cluster B-This group is having policy holders with claims of higher values. Claim were Dominated by eye and adnexa problems. This group is having policy holders of early forties and their claims are being dominated by miscellaneous expenses and also this groups dominated by individual health insurance policy holder.

Cluster C- This group has low risk customers whose claim values are lowest amongst their group. Most of the claims pertain to cases diagnosed with pain abdomen problems of medium age groups with surgical less treatment. The treatment value in all headers is low compared to other groups other than pre-hospitalization expenses.

Cluster D- This group is dominated by group floater policies. Due to minimum sum assured, total claim is second lowest amongst the group. Most of the claims are surgically in nature and the diagnosed claims are of mostly normal pregnancy related with moderate surgical expenses. The claim value is having least post and pre_hospitalization expenses.

Thus corresponding confusion matrix for the clustering model based on policy type attribute is given in Table I.

```
              -: Evaluation And Modeling On The Training Set:-

   •  Clustered  Instances
   0   3699 (25%)
   1   7109 (47%)
   2   923   (6%)
   3   3269 (22%)

Class Attribute: Txt_Type_Of_Policy

Classes to Clusters:

0       1       2       3 ← assigned to cluster

1359     2534   52      733     |       1
82      265     60      81      |       2
768     1289    62      433     |       3
1490    3021    749     2022    |       4


Cluster A ← 3
Cluster B ← 1
Cluster C ← 2
Cluster D ← 4
```

From the confusion matrix given in the table 1, it is evident that cluster 1 has policy holders mostly dominated by individual policies,follwed by cluster C,cluster A and Cluster D having policy type prefrences of individual,group and group floater policies. The tuples of clusters B with major contributions of 47% of instances follwed by Cluster A,Cluster D,and Cluster C.By consedering the above facts,each cluster is assigned ranksas cluster B ahving rank 1 followed by cluster A,Cluster D and Cluster C having ranks 2,3 and 4 respectivley.

B. Observation Based On Association Rule:-

Prediction of claim levels are based on distribution of diseases among the instances of the working set. Apriori assciation rule performed on claim dataset gives the association among attributes in the claims dataset.Different association Rule express different regularities that underlie the dataset and predict different things. The values for number of rules considerd,the decrease for minimum support (delta factor) and minimum confidence values are 3, 0.95 and 0.9 respectivley. The number of itemset and correspoding rule mapping is given in table II.

	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Rule Set Size	9	36	84	126	126	84	36	9	1

Table II Aprioi Association Rule Set On Claim Dataset

Best three rules genrated with 14490,14986 and 14983 instances associated and its description as follows:

First rule specifies it is of single rule set category and if miscellaneous expenses of range(0-285712)has the impact on post hosptalization expenses of range (0-297650) with cinfidence level of 1. In addition,lift = 1 indicates that the having miscellaneous expenses within the range increases the probabilty of producing the post-hospitalization expenses by factor of 1. Levearge is the propotion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other.

In the second rule,which is of two set category, the attribute invertigation charges and miscellaneous charges have significant impact on post hopitalization expenses of the claim. The ranges of [0-153538.66] for inverstigation charges and [0-2857]for miscellaneous have the impact on post hospital expenses within the rang [0-297650.66] with the confidence level of 1.

Third rule which is of two set category, that predicts post hospitalization expenses.It is obsevred from the third rule that if surgery charges of range(0 to 116666.67) and miscellaneous charges of range (0 to 285712) influences the post hospitalization expenses to fall within the range 0 to 297650.67.

The component lift=1 indicates that having LHS predicates increases the probability of RHS predicates by a factor of 1 in the best 10 rule sets.

As a reference to third rule the component post hospitalization of claim fall in the range (0-297650.67) is dependent on surgery charges of range (0-116666.67)and Miscellaneour charge (0-285712).

Similarly,other two item sets rule 4,rule 5 and rul 9 justifies the dependent factors(LHS) on RHS attribute Post Hospitalisation Expenses. Sixth rule is of three item set category specifies if the claim amount lies within the range 0 to 456470.67,investigation charges charges between 0 to 0-153538.67 and Miscellaneous charges between 0 yo 285712 has an significant on post hospitalisation expenses of range 0 to 297650.67. This rule has 14979 instances justifies that Similarly, rule numbers 7,8 and 10 are of three item category justifies the dependent factors (LHS) on RHS attribute post hospitalization expenses.

V. CONCLUSION

Data mining provides the technology to analyze mass volume of data and/or detect hidden patterns in data to convert raw data into valuable information. Clustering is technique applied on health insurance claim dataset to segment the customers. Clusters revealed the preferences of customers towards the products and factors that influence total claim. Association rule applied on working dataset to predict nature of the claim. The process of combining segmentation with data mining provides marketers with high quality of information on how their customers shop for and purchase their products or services. By combining the standard market with data mining techniques we can predict and model behavior of segments. Although the paper mainly focuses on the insurance industry, the issues and applications discussed are applicable to other industries, such as banking industry, retail industry, manufacture industries, and so on.

REFERENCES

[1] Pramod Prasad, Latesh G. Malik.(2011)Generating customer profiles for retail stores using clustering techniques,International Journal On Computer Science And Engineering, Vol.3,pp.2506-2510.

[2] Larose,DT(2006)Data Minings method and Models,Hoboken,New-Jersy,John Wiley & sons,Inc.

[3] Roosvelt Mosley,(2005),The Use Of Predictive Modeling in the Insurance Industry,Pinnacle actuarial resources.

[4] Westphal.Cand T.Blaxton (2005),Introduction To Data Mining and Knowledge Discovery,Two Cows Corporation,Third Edition.

[5] Kamber.M.J.Han.(2008),Data Mining : Concept And Techniques,Morgan Kaufman.

[6] Huang.S.C.Chang,&H..H. Wu(2009),A case study of applying Data Mining in outfitters's customer value analysis ,Expert Systemwith Applcations,Vol.36 Issue6,pp5909-5915.

[7] Blocker, C.P., &flint, D.J.(2007).Customr segments as moving targets:Integarting Customer Value Dynamism Into Segment Instability Logic.Industrial Marketing Management.Vol.36,pp.810-822.

[8] Kanwal Garg,Dharminder Kumar and M.C. Garg,(2008),Data Mining Techniques for Identifying The Customer Behaviourof Investment in life insurance sector india, International Journal Of Information technology and Knowledge Management,Vol.1 ,Issue.1,pp.51-56.

[9] Tsiptsis.K and A. Chorianopoulos , Data Mining Techniques In CRM ,Inside Customer Segmentation (2009).Jhon Wiley &Sons..Ltd.Second Edition.

[10] Two Crows Corporation(1998)Introduction To Data Mining And knolwedge Discovery,Second Edition Patomac,MD.

[11] Pieter Adriaans,Dolf Zantinge,(2011),Data Mining,Pearson Education Ltd,Sixth Impression,2011.

[12] Migueis,V.l.Camanho,A.S.Joao Falcao e Cunha (2012),Customer Data Mining for LifeStyle Segmentation,Expert Systems eith applications,Vol.39.3959-9366.

[13] Hosseini,M.,Anahita,M.Mohammad,R,G.(2010).Cluster Analysis using data mining approach to develop CRM methodology to assess the customer loyalty.Expert Systems With Applications,Vol.37.pp.5259-5264.

[14] Samira Malekmohammmadi Golsefid,Mehdi ghazanfari, Somayeh Alizadeh(2007),Customer Segmentation in Foreign Trade based on Clustering Algorithm,World Academy of Science,Engineering and Technology Vol.28,pp.405-4011.

[15] An integrated data minig and behavioral scoring model for analyzing bank customer, Nan-Chen Hsieh,Expert System With Applications 27 (2004),623-633.

[16] Knowledge Management In CRM using Data Mining Tchniques,Sunil yadav,Aaditya Desai,Vandana Yadav,International Journal Of Scientific & Engineering Research, Volume 4,Issue &, July-2013.

[17] The Effect Of Clustering in Apriori Data Mining Algorithm : A Case Study, Nergis Yilmaz and Gulfem Isiklar Alptekin, Proceedings of the World Congress on Engineering 2013 Vol III,WCE 2013,July 3-5,2013,London,U.K.

[18] Data Mining: An Overview from a Database Perspective Ming-Syan Chen,Senior Member,IEEE,Jiawei Han,Senior Member,IEEE,and Philip S. Yu,Fellow,IEEE, IEEE transaction On Knowledge And Data Engineering,Vol.8,No.8,No. 6,December 1996 .

[19] Mining Changes in customer behaviour in retail marketing,Mu-Chen Chen,Ai-Lun Chiu,Hsu-Hwa Chang,Expert System with Applications 28,(2005), 773-781.

[20] Data Mining By Heidi Kuzma And Sheila Vaidya.